2021年6月18日　統計数理研究所　オープンハウス

データ多様化時代の統計的データ結合技術

山下　智志　　データ科学研究系　教授

【社会的背景】

ビジネスにおいて企業データや個人データの入手機会が増え、またデジタルストレージ価格の低下から、データを大量に保存するようになった。

近年，インターネット上の情報，公的統計ミクロデータ，民間企業のデータなどの様々なデータが利用可能になっており，これらのデータを何らかの形で結合することができれば，新たに統計調査やデータの収集等を行うことなく，情報量（変数）の多い有用なデータを構築することが可能となる．

このような状況の中，複数のデータをレコード単位で結合するデータリンケージ（Data Linkage）の手法が，様々な分野で注目を集めている．例えば，企業の過去のデータを基にデフォルト確率予測モデルを構築し，信用力の評価を行う場合には，企業のデフォルトに関する大規模なデータが必要となるが，その際に多様な性質を持つ複数のデータを結合することにより，様々な財務指標や企業の属性情報などの分析に利用可能な情報を効率的に増加させることが可能となり，データを収集する際のコストの削減が期待される．

公的統計ミクロデータに関しては，昨今，公的統計ミクロデータ研究コンソーシアムの設立や公的統計のオーダーメード集計の利用条件等の緩和など，その利活用に向けた機運が急速に高まっている．また，平成29年に決定された「統計改革推進会議最終取りまとめ」（平成29年5月19日統計改革推進会議決定）や，統計委員会の答申を受けて平成30年に閣議決定された第III期「公的統計の整備に関する基本的な計画」（平成30年3月6日閣議決定）では，今後，企業の保有するビッグデータなどの公的統計への活用について検討を進めることとされている．これらの決定等を踏まえ，政府は平成 30 年3月6日に，公的統計ミクロデータの更なる利活用を含む「統計法及び独立行政法人統計センター法の一部を改正する法律案」（閣法第 34 号）を国会に提出し，平成 30 年5月 25日に可決・成立した（平成 30 年法律第 34 号）．政府統計を取り巻くこのような状況を鑑みれば，公的統計ミクロデータと企業の保有する様々なデータとのデータリンケージや統計的マッチングは，既存のデータを有効に活用した有用なデータの構築につながるものであり，今後一層，重要な研究課題になると考えられる．

【データ結合とは何か】

複数のデータを結合する際に，各レコードを識別できる照合キー（共通一連番号，名称，所在地など）が存在する場合には，それらを利用してレコードを結合する完全照合（Exact Matching）を行うことが可能である．しかし，異なる機関が整備する企業データに関しては，秘匿性の観点から名称や所在地などの個体を特定できる情報を相互に利用することができず，資本金や売上高などの限られた情報のみが利用可能である場合が多いと想定される．そのような場合には，複数のデータに共通に含まれる変数を基に，何らかの意味で類似したレコード同士を結合する方法が用いられる．これを統計的マッチング（Statistical Matching）という

多数のデータベースから情報を効率的に抽出する技術に関する問題

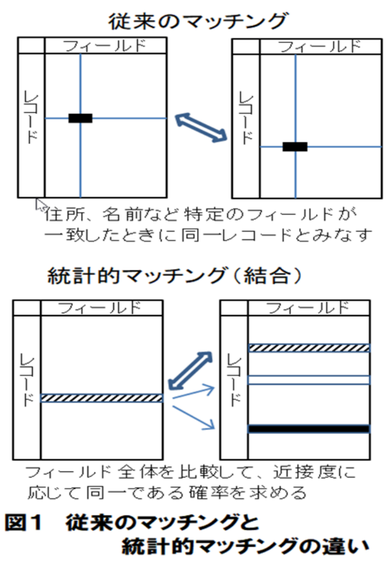


図1　従来のマッチングと統計的マッチングの違い

既存の統計モデルは単一のデータベースが前提となっており、複数の性質が異なるデータベースが存在することを前提とした理論ではない。そのためデータベースを結合し単一データベースしなければほとんどの統計的手法が適用できない。この際以下のような問題が生じる。

【問1】
統計論的マッチングによるデータ統合が可能か

「名前や住所が同じならば同じ個人（企業）」という決定論的マッチングから、フィールド全体を利用し同じ傾向にあるデータだから同じ個人（企業）と判断する統計論的マッチングへ

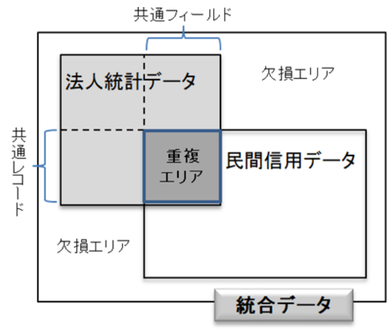


図2　統合データベース作成のイメージ

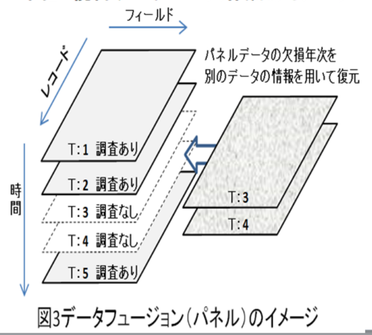


図3データフュージョン（パネル）のイメージ

【問2】
一部の情報を共有するデータベースの結合方法は
レコードとフィールドの一部を共有する場合、欠損エリアと重複エリアが存在し、それに対するバイアスが発生しない処理技術が必要となる。

【問3】
質と量が極端に異なるデータに有効な統計手法は
低質大量データと高質少量データを結合し、統計モデルの作成する方法論と利用可能性について検討する

【問4】パネルデータに対するデータフュージョンとは
調査時期によって調査対象が異なる、データ欠損時期があるなどの不完全パネルデータを復元するための統計処理

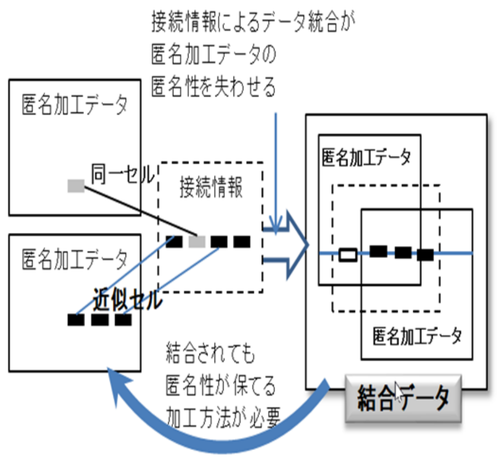


図4　プロファイリングによる非匿名化

【問５】
データベースの結合を前提とした匿名加工のあり方
個人データの結合（プロファイリング）をすることによって、もともと匿名加工されていたデータの匿名性が失われることがある。逆に、データ結合されることを前提とした匿名加工の方法が必要とされる。

【問6】
政府ミクロデータと民間データの結合により新たな知見を得られるか
問１、問２、問４の結果を政府ミクロデータと民間データの結合に応用する。特にセンサスを前提とした政府データのフィールドを民間データで補間し情報精度を向上させ、民間データをセンサスデータで非観測レコードを保管する。

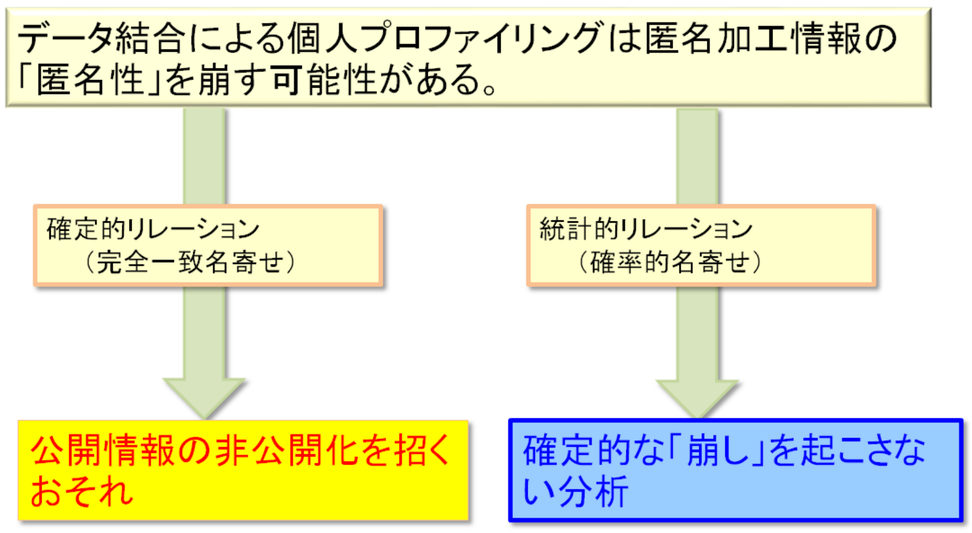
【ウエイト付き距離を用いた多項ロジットモデル（Multinomial Logit Model）に基づく新たな統計的マッチングの手法】

提案手法により，利用可能な変数が少なく，名称，所在地などの詳細な文字情報がない企業データに対しても，効率的かつ効果的な統計的マッチングを行うことが可能となる．また，本研究で提案する統計的マッチングのモデルにより，距離のウエイトを最尤法の枠組みで統計的に（最尤法により）推定することが可能となり，これまで過去の経験や専門的な知識に基づいて設定されることが多かった距離のウエイトについて，データに基づき最適な値を推定することができる．さらに，マッチングの正しさに関する確率（マッチング確率）を推定することが可能となり，マッチングの精度の定量的な比較を行うことができる．

なお，名称，所在地などの詳細な文字情報に基づく統計的マッチングでは，同一の対象に対する複数の表現（漢字，平仮名，片仮名，アルファベット等）が存在する標記ゆれの問題があり，これがマッチングを困難なものとしているが，距離に基づく統計的マッチングではそれらの文字情報を用いないため，そのような表記ゆれの問題は生じない．

また，詳細な文字情報によるマッチングは個別のレコードの特定につながるおそれがあるが，提案手法ではマッチング確率を算出するのみであり，直接的な対象の特定を行っているわけではない．提案手法を実際のデータ（平成24年経済センサス--活動調査のミクロデータ及び帝国データバンクのデータ）に適用した結果，多項ロジットモデルは適切に推定されており，最も当てはまりの良いウエイト付き絶対値距離の対数変換を用いたモデルに基づく統計的マッチングは，マッチングの正解率の観点から，従来の研究で用いられている最近隣法（Nearest Neighbor Method）よりも優れていることが示された．

データ結合と匿名加工情報の関係



【提案した統計的マッチングの手法の速度と精度の改良】

<u>主成分分析を層化による計算時間の短縮効果</u>

提案手法が優れた性能を発揮することが示されたものの，レコード間の距離や対数尤度の計算に伴う計算量の問題は依然として残っている．
例えば，経済センサスのようにサイズの大きなデータを扱う場合には，距離や対数尤度の計算の対象となるレコードの組合せの数も非常に多くなることから，多項ロジットモデルの推定やレコードのマッチングに相当な時間がかかるものと考えられる．

このような問題に対して，第3章では，主成分分析の結果（第1主成分得点）に基づいてデータを層化し，同一又は近隣の層のレコードのみを距離や対数尤度の計算の対象とすることによって計算の効率化を図り，マッチングの精度を大きく低下させない形で計算速度を向上させる方法について検討する．

提案手法を経済センサスミクロデータ及び帝国データバンクデータに適用した結果，層の数を適切に設定することにより，正解率の低下を最小限に抑えつつ，計算時間を大幅に削減できることが示された．

表 4.7　トップのレコードに対する正解率

	地域 A	地域 B	地域 C
多項ロジットモデルに基づく方法			
ウエイト付き Euclid 距離 (2 乗)	0.346	0.463	0.243
ウエイト付き絶対値距離	0.407	0.528	0.350
ウエイト付き絶対値距離 (対数変換)	0.449	0.582	0.390
最近隣法			
Mahalanobis 距離	0.021	0.092	0.030
Gower 距離	0.209	0.367	0.220
fastLink	0.090	0.278	0.104

チューニングパラメータの置き方の違い？　転移性・計算負荷を確保して精度を犠牲にした？

$$\begin{aligned}\xi_{ij} &= \Pr(M_{ij}=1 \mid \boldsymbol{\delta}(i,j),\boldsymbol{\gamma}(i,j)) \\ &= \frac{\lambda \prod_{k=1}^{K}\left(\prod_{\ell=0}^{L_k-1}\pi_{k1\ell}^{\mathbf{1}\{\gamma_k(i,j)=\ell\}}\right)^{1-\delta_k(i,j)}}{\sum_{m=0}^{1}\lambda^m(1-\lambda)^{1-m}\prod_{k=1}^{K}\left(\prod_{\ell=0}^{L_k-1}\pi_{km\ell}^{\mathbf{1}\{\gamma_k(i,j)=\ell\}}\right)^{1-\delta_k(i,j)}} \end{aligned} \quad (3)$$

<u>制約付きマッチングによる精度向上</u>

統計的マッチングの手法を単純に適用した場合，レコードが複数回使用されることに関する制約を設けていないため，1つのレコードに複数のレコードがマッチングされる可能性があり，そのような場合，正しいマッチングが実現できず，マッチングの精度の低下につながるおそれがある．

そこで，多項ロジットモデルにより推定されたマッチング確率を用いて，統計的マッチングの問題を重み付き2部グラフ（Weighted Bipartite Graph）の最適マッチングの問題として定式化した上で，この問題に対する効率的なアルゴリズムであるハンガリー法（Hungarian Method）を適用することにより，1対1の制約付きマッチング（Constrained Matching）を実現しつつ，更なるマッチング精度の向上を図っている．ハンガリー法のアルゴリズムは実装しやすく，その計算速度は速い．

提案手法を複数の地域のデータに対して適用した結果，多項ロジットモデルに基づく統計的マッチングの方法を単純に適用した場合と比較して，全ての地域において統計的マッチングの正解率が向上することが確認できた．